On implied genetic effects, relationships and alternate allele coding

Bruce Tier, Karin Meyer & Andrew A. Swan

Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia

Summary

This paper examines some of the implied effects commonly assumed when building relationship matrices. We propose the inclusion of an additional 'individual' in the genomic relationship matrix which models the mean of the founder population. It is shown that this resolves the problem of inconsistent prediction error variances due to alternate allele coding schemes.

Keywords: Allele coding, implied genetic effects, prediction error variance

Introduction

Together with the genetic variance, relationships between animals are essential to model covariances among individuals when estimating genetic parameters or predicting breeding values (EBVs). In the absence of genomic information, relationships are based on co-ancestry (identity by descent or IBD) using the numerator relationship matrix (**A**). High density tests for single nucleotide polymorphisms (SNPs) have been available for many years and now young animals are tested routinely for tens of thousands of loci. Commonly, these data are 0, 1 or 2, the number of copies of one of two possible alleles at a given locus. They can be used to build a genomic relationship matrix (**G**) which is based on identity by state (IBS). Important parameters required when building **G** are the allele frequencies ($\mathbf{p}_F$) for each locus in the founder population.

To their surprise, Forni *et al.* (2011) found that applying alternate allele codings did not necessarily affect relative EBVs. Strandén & Christensen (2011) further showed that, other than a change in mean, coding of alleles had no effect on EBVs but did affect prediction error variances (PEVs) and accuracies. Tier *et al.* (2015) demonstrated that there are infinite variety of **G** matrices that provide the 'same' EBVs.

This paper illustrates the implied genetic group in **A** and the implied founder in **G** which are generally disregarded. We show that the latter can explain a number of observed phenomena and that inclusion of the implied founder results in consistent PEVs and accuracies.

The implied genetic group in A

We generally assume that founder animals are a random sample of a huge population that has been mating randomly and has an infinite number of loci. The consequence of this is that the part of **A** that corresponds to founders is modeled with an identity matrix. This says that each founder has two unique gametes that are not shared by any other animals (they are not inbred and the covariances between founders is zero). Having renumbered animals from 1 to n such that parents have lower numbers than their progeny, **A** with elements a_{ij} can be built one row (column) at a time as $a_{ii} = 1 + 0.5a_{sd}$ and $a_{ji} = 0.5\left(a_{js} + a_{jd}\right) = a_{ij}$ for $i > j$ and s and d the parents of animal i.

The inverse $\mathbf{A}^{-1}$ is required when estimating genetic parameters or predicting breeding values. Following Henderson (1976) it is commonly built directly, one animal at a time, by accumulating contributions to elements a^{kl} (of $\mathbf{A}^{-1}$) relating to itself (i), its sire (s) and its dam

(d): These contributions are

$$\begin{array}{llll} m_i/4 & \text{to} & a^{ss}, a^{sd}, a^{ds} \text{ and } a^{dd} & \text{(parent - parent)} \\ -m_i/2 & \text{to} & a^{is}, a^{id}, a^{si} \text{ and } a^{di} & \text{(parent - individual)} \\ m_i & \text{to} & a^{ii} \quad \text{(individual)} \quad \text{where} & m_i = (1 - 0.25(a_{ss} + a_{dd}))^{-1} \end{array} \tag{1}$$

In the original formulation, when one or both parents are unknown the corresponding terms are omitted in (1). However, by considering unknown parents to be 'phantom' animals (Westell *et al.*, 1988) with the number 0, all contributions could be made by augmenting $\mathbf{A}^{-1}$ with a zeroeth row and column. This equation can be, and usually is, omitted and consequently its solution in any analysis is zero. This zeroeth row represents the implied genetic group in **A**. Consequently each animal's EBV can be considered as the sum of a genetic group effect (of zero) and it's deviation from that group effect, as in a model with a single genetic group included.

The implied founder animal in G

Consider the linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \mathbf{e}$, where the data ($\mathbf{y}$) are a function of fixed ($\mathbf{b}$) and random animal ($\mathbf{u}$) effects and a residual ($\mathbf{e}$). **X** and **W** denote the pertaining incidence matrices. Random effects **u** and **e** are assumed to have null means and variances $\sigma_g^2\mathbf{G}$ and $\sigma_e^2\mathbf{I}_n$, respectively, with **G** as given below (3) and $\mathbf{I}_n$ an identity matrix of size n. Mixed model equations (MME) are

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{G}^{-1}\sigma_e^2/\sigma_g^2 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{pmatrix} \tag{2}$$

Let **C** denote the coefficient matrix in (2). PEVs are proportional to the diagonal elements (c^{ii}) of $\mathbf{C}^{-1}$. The accuracy for animal i is then $r_i = \sqrt{1 - c^{ii}\sigma_e^2/(g_{ii}\sigma_g^2)}$, with g_{ii} the i–th diagonal element of **G**.

The genomic relationship matrix is generally built using VanRaden's (2008) first method:

$$\mathbf{G} = (\mathbf{Z} - 2\mathbf{P})(\mathbf{Z} - 2\mathbf{P})'/t \tag{3}$$

where **Z** is an (n) animal by (g) genotype matrix with elements 0, 1 or 2 which represent the number of copies of one of the two alleles at each animal-locus element, $\mathbf{P} = \mathbf{1}_n\mathbf{p}'$ is a matrix with rows $\mathbf{p}'$, the vector of allele frequencies and $\mathbf{1}_n$ a unity vector of size n with the scaling factor t, commonly calculated as $2\mathbf{p}'(\mathbf{1}_n - \mathbf{p})$. Conceptually, vector **p** is equal to $\mathbf{p}_F$ – the frequencies in the founder population – but it is usually derived from the observations. However, any vector can be used to form **P** in the numerator of (3): choice of **p** affects how allele counts are centered which can be thought of as equivalent to altering allele coding. For instance, using $\mathbf{p} = 0.5$ changes coding from 0, 1 and 2 to −1, 0 and 1 which can be computationally advantageous.

We can augment **G** with an implied founder, analogous to the implied genetic group for **A**, by adding an animal (treated as animal "0" or A0 from here on), which models the mean of the founder population. This involves adding a row in **Z** with values equal to twice the (assumed) allele frequencies for $\mathbf{p}_F$. The resulting EBVs depend on the choice of **p** to form **P** – for $\mathbf{p} = \mathbf{p}_F$, its solution is zero (and **G** is not positive definite as elements pertaining to A0 are 0). For $\mathbf{p} \neq \mathbf{p}_F$, we can adjust the EBVs of (actual) animals by subtracting the mean of the founders, yielding $\hat{v}_i = \hat{u}_i - \hat{u}_0$. In matrix terms

$$\hat{\mathbf{v}} = \left[-\mathbf{1}_n \vdots \mathbf{I}_n\right]\hat{\mathbf{u}}^\star = \mathbf{Q}\,\hat{\mathbf{u}}^\star \qquad \text{with} \qquad \hat{\mathbf{u}}^{\star\prime} = \left[\hat{u}_o \vdots \hat{\mathbf{u}}'\right] \tag{4}$$

Let subscripts 'F' and '$\neq F$' denote matrices constructed using $\mathbf{p} = \mathbf{p}_F$ and $\mathbf{p} \neq \mathbf{p}_F$, respectively. Matrix $\mathbf{Q}$ provides the transformation from $\mathbf{G}_{\neq F}$ to $\mathbf{G}_F$ and the part of $\mathbf{C}^{-1}$ required to obtain PEVs for (actual) animals.

$$\mathbf{G}_F = \mathbf{Q}\,\mathbf{G}_{\neq F}\mathbf{Q}' \qquad \text{and} \qquad \mathbf{C}^u_F = \mathbf{Q}\,\mathbf{C}^u_{\neq F}\mathbf{Q}' \tag{5}$$

where $\mathbf{C}^u_F$ denotes the submatrix of $\mathbf{C}^{-1}$ for animals 1 to n. Thus the anomaly of alternate allele codings for PEVs is resolved.

Using a small example comprised of four animals (denoted as A1 to A4) we illustrate how inclusion of this extra animal results in a single set of PEVs for any allele coding scheme. We fit a model containing an overall mean as the only fixed effect and genetic effect for each animal in $\mathbf{u}$. Allele frequencies in the founder population are assumed to be $\mathbf{p}_F = 0.5$. Geno- and phenotypes for A1 to A4, are given in Table 1. We consider three cases (with $t = 7$ and $\sigma^2_e = \sigma^2_g$ throughout):

I $\mathbf{G}$ is constructed with $\mathbf{p} = 0.5$ for all loci, so that $-2\mathbf{P}$ changes the allele coding to $-1, 0, 1$.

II As I but assuming $\mathbf{p} = 0$ for all loci, i.e. genotypes are processed uncentered.

III As II, but adding an implied founder, A0, to our population and model. Elements in the pertaining row in $\mathbf{Z}$ have values equal to twice the allele frequency in the founder population. Since we assume $\mathbf{p}_F = 0.5$ for all loci, all elements in the row are unity.

Matrices $\mathbf{G}$, $\mathbf{C}$ and $\mathbf{C}^{-1}$ and solutions to the MME and their calculated accuracies are summarised in Tables 2 and 3. Results illustrate the shift in EBVs due to a change in $\mathbf{p}$ used to calculate $\mathbf{G}$ while their relative differences remain the same. As emphasized by Strandén & Christensen (2011), PEVs and the resulting accuracies differ considerably. For the same $\mathbf{p}$ (cases II and III), inclusion of the implied founder animal had no effect on the matrix elements in $\mathbf{G}$ and $\mathbf{C}^{-1}$ pertaining to animals 1 to 4 nor the respective solutions. Adjusting EBVs for the implied founder's solution of 0.154 yields transformed values equal to those obtained in case I.

The extra animal in $\mathbf{Z}$ (with genotypes $2\mathbf{p}_F$) in case III represents the founder population and forms the basis of the relationships in $\mathbf{G}$. Its (non-zero) solution results from $\mathbf{p} \neq \mathbf{p}_F$ and is the difference in mean between the founder population and the population assumed with $\mathbf{p}$. Without this animal the mean of the founder population represented in $\mathbf{P}$ is assumed to be zero. When $\mathbf{p} = \mathbf{p}_F$ all is well, but we generally cannot know $\mathbf{p}_F$. Cases I and II show that alternate vectors $\mathbf{p}$ do not matter if all we are interested in is $\hat{\mathbf{u}}$. When we need accuracies, however, this assumption is crucial. Case III shows that it is not the allele codings, but the different assumed $\mathbf{p}$ that are the problem. Obvious choices are 0.5, the observed values or the imputed frequencies among pedigree founders. The first implies maximum variance of SNPs, the third aligns $\mathbf{G}$ and $\mathbf{A}$. Case III also shows that the appropriate prediction error variances for any assumed $\mathbf{p}$ are retrieved. Equivalent SNP based models can produce the same EBVs and prediction error covariances. These can be transformed in the same way by including an extra animal with genotypes $2\mathbf{p}_F$.

Considerable effort has been made to resolve problems associated with combining relationships based on IBD and IBS. These problems arise because the different matrices relate to different founder populations with different means, both assumed, explicitly or implicitly, to be zero. While the founders in $\mathbf{A}$ are known and the founders in $\mathbf{G}$ are their unknown ancestors it is unlikely that they share the same mean. The use of meta-founders (Legarra *et al.*, 2015) to modify relationships among pedigree founders, so that they are all related, is a way of projecting the founder population beyond the pedigree founders altering both their mean and the additive genetic variance. If $\mathbf{p}_F$ was known the implied founder would fulfil a similar function to a meta-founder (Legarra *et al.*, 2015), and possibly one founder could serve both purposes when animals with and without genotypes are being analysed together. We can be fairly certain that the observed $\mathbf{p}$ is not $\mathbf{p}_F$. While the choice of $\mathbf{p}$ used when building $\mathbf{G}$ has no effect on the results, this may not be the case when **G** is combined with **A**; however, this is beyond the scope of this paper. Strandén & Christensen (2011, Eqn. 2) showed the mean as made up of two components – an overall mean plus the genetic mean of the founder population. They applied the joint mean as the denominator in the accuracy equation, which has infinite variance. Only one component – that corresponding to the mean of the founder population should be used – and when it is, the denominator in the accuracy does not have infinite variance. When developing a SNP based model for single step evaluation, Fernando *et al.* (2014) imputed the marker covariates for ungenotyped individuals and added an equation to represent the mean of a random animal drawn from the unselected founder population. This parallels the function of the implied founder and enabled modeling the residual imputation effects with a null mean.

In practice, these are problems in the extensive livestock industries. Unlike intensive industries such as dairy and poultry where genotyping animals is widespread and systematic, this process is more haphazard in beef and sheep where there are still large sections of the population without genotypes. Often problems are exacerbated by the presence of multiple breeds and crossbreds, all with different founder populations. Use of meta-founders to model relationships among founders of different breeds can accommodate relationships within and between breeds. It is likely that separate implied founders will need to be included for each breed.

Conclusions

Analyses using the genomic relationship matrix invoke multiple assumptions which can be confusing and are sometimes inconsistent with other assumptions. We illustrate that particular care needs to be taken in modeling means of founder populations, and propose identification of an implied founder that may help resolve some of these problems.

List of References

Fernando, R. L., J. C. M. Dekkers & D. J. Garrick, 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Forni, S., I. Aguilar & I. Misztal, 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43(1).

Henderson, C. R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar & I. Misztal, 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200(2):455–468.

Strandén, I. & O. F. Christensen, 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43(1):1–11.

Tier, B., K. Meyer & M. H. Ferdosi, 2015. Which genomic relationship matrix? Proc. Ass. Advan. Anim. Breed. Genet. 21: Paper no. 83.

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11):4414–4423.

Westell, R. A., R. L. Quaas & L. D. Van Vleck, 1988. Genetic groups in an animal model. J. Dairy Sci. 71(5):1310–1318.

Acknowledgements

AGBU is a joint venture of NSW Department of Primary Industries and the University of New England. Work was supported by Meat and Livestock Australia grant L.GEN.1704.

Table 1. Geno- and phenotypes for numerical example.

Animal	Allele counts	Record
A1	1 1 0 1 1 1 0 0 1 2 1 1 0 1	5
A2	0 1 1 2 0 1 1 1 1 1 2 1 1 2	11
A3	2 0 0 1 1 0 0 0 1 2 1 2 1 2	7
A4	1 0 0 1 1 0 0 1 0 1 1 2 2 2	17

Table 2. Genomic relationship matrices **G** *for numerical example.*

	Case I (**p=0.5**)				Case II (**p=0**)				Case III (**p=0** with A0[1])				
	A1	A2	A3	A4	A1	A2	A3	A4	A0	A1	A2	A3	A4
A0									2.00	1.57	2.14	1.86	1.71
A1	0.71	0.00	0.57	0.14	1.86	1.71	2.00	1.43	1.57	1.86	1.71	2.00	1.43
A2	0.00	0.71	0.00	0.14	1.71	3.00	2.00	2.00	2.14	1.71	3.00	2.00	2.00
A3	0.57	0.00	1.29	0.86	2.00	2.00	3.00	2.43	1.86	2.00	2.00	3.00	2.43
A4	0.14	0.14	0.86	1.14	1.43	2.00	2.43	2.57	1.71	1.43	2.00	2.43	2.57

[1] Implied founder

Table 3. Coefficient matrices and solutions for numerical example.

Case I (**p=0.5**)					Case II (**p=0**)					Case III (**p=0**, with A0[1])					
$\hat{b}$	A1	A2	A3	A4	$\hat{b}$	A1	A2	A3	A4	$\hat{b}$	A0	A1	A2	A3	A4
Coefficient matrix **C**															
4.00	1.00	1.00	1.00	1.00	4.00	1.00	1.00	1.00	1.00	4.00	0.00	1.00	1.00	1.00	1.00
										0.00	3.32	-1.48	-1.27	0.45	-0.82
1.00	3.81	-0.25	-2.07	1.23	1.00	4.47	-1.22	-3.01	1.86	1.00	-1.48	5.13	-0.65	-3.21	2.22
1.00	-0.25	2.50	0.43	-0.48	1.00	-1.22	2.13	0.94	-1.09	1.00	-1.27	-0.65	2.62	0.77	-0.78
1.00	-2.07	0.43	4.12	-2.14	1.00	-3.01	0.94	5.04	-2.88	1.00	0.45	-3.21	0.77	5.10	-2.99
1.00	1.23	-0.48	-2.14	3.38	1.00	1.86	-1.09	-2.88	3.92	1.00	-0.82	2.22	-0.78	-2.99	4.13
Inverse of coefficient matrix: $\mathbf{C}^{-1}$															
0.64	-0.33	-0.28	-0.52	-0.44	2.20	-1.71	-2.03	-2.17	-1.90	2.20	-1.72	-1.71	-2.03	-2.17	-1.90
										-1.72	1.88	1.55	1.91	1.81	1.63
-0.33	0.53	0.15	0.43	0.20	-1.71	1.73	1.72	1.90	1.49	-1.71	1.55	1.73	1.72	1.90	1.49
-0.28	0.15	0.54	0.21	0.24	-2.03	1.72	2.47	2.05	1.89	-2.03	1.91	1.72	2.47	2.05	1.89
-0.52	0.43	0.21	0.85	0.56	-2.17	1.90	2.05	2.59	2.12	-2.17	1.81	1.90	2.05	2.59	2.12
-0.44	0.20	0.24	0.56	0.74	-1.90	1.49	1.89	2.12	2.12	-1.90	1.63	1.49	1.89	2.12	2.12
Solutions[2]															
9.84	-2.41	0.85	-0.55	2.77	9.68	-2.26	1.01	-0.40	2.92	9.68	0.15	-2.26	1.01	-0.40	2.92
												-2.41	0.85	-0.55	2.77
Accuracies[3]															
	0.51	0.50	0.58	0.59		0.26	0.42	0.37	0.42		0.24	0.26	0.42	0.37	0.42
												0.51	0.50	0.58	0.59

[1] Implied founder

[2] First line: solutions, second line: animal solutions for case III adjusted for A0

[3] First line: from MME, second line: adjusted